NANYANG TECHNOLOGICAL UNIVERSITY
Programming Interview

Spatially Resolved Tumor Purity Report

Ruisi GENG

December 8, 2021

1 Executive Summary

The article developed an advanced deep multiple instance learning model to accurately predict tumor purity from HE stained digital histopathology slides, and the model was applied to obtain prediction results based on two TCGA cohorts, one slide from each The model was applied to obtain predictions based on two TCGA cohorts, one from fresh-frozen sections and the other from formalin-fixed trafficking-embedded sections, where the final results were highly consistent with genomic tumor purity values. Another achievement obtained obtained a series of spatially resolved tumor purity maps presented with an illustration of spatially variation in a sample. With a throughout estimation model architecture, pathologists can therefore choosing sample to minimize normal cell contamination in high throughput genomic analysis. also greatly improves the cost efficiency by reducing the effect of variation in internal variables on the results.[1]

2 Introduction

Tumor purity, the proportion of cancer cells, accounts for the tumor tissue. It is widely accepted that accurate tumor purity estimation is of great clinical importance, especially in pathologic evaluation and minize the normal cell contamination in high throughput genomic analysis [1]. Therefore, it is of great importance to give an accurate tumor purity estimation. The paper introduced two popular application methods and one novel ML model: percent tumor nuclei estimation and genomic tumor purity inference. However, both of the previous estimations have pros and cons. With a detailed introduction, the multiple instance learning (MIL) model predicts sample-level tumor purity from H E stained digital histopathology slides.

MIL was first proposed by Dietterich et al.[2] in Drug Activity Prediction. Most drugs are composed of small molecules that can bind to large proteins. One of the low-energy shapes that can be made into a drug can be tightly bound to the target site. A molecule can have more than one low-energy shape, but at the time, researchers could only tell if a molecule could be made into a drug, not which low-energy shape of the molecule was at work. If a simple classification algorithm was used, the low-energy shape of all molecules that could make a drug was treated as a positive example, and the opposite was treated as a negative example[3]. A pharmaceutical molecule may have hundreds of low-energy shapes, but only one of them is really working, so he proposed the concept of MIL. The molecule that can make drugs is naturally a positive bag, and the opposite is a negative bag. Until recent decades, the existing MIL algorithms could be classified into two groups: bag space algorithms and instance space algorithms.[4]

3 Materials & Methods

In essence, the unit in data of MIL's dataset is a bag. Take binary classification as an example. A bag contains multiple instances. If all instances are marked as negative, then this bag is negative, and vice versa. Let Y be the label of package $X{=}x_1, x_2, ..., x_n$, and each example xi corresponds to a label y_i, then the label of the package can be expressed as

$$f_x(x) = \begin{cases} +1, \text{if} \forall y_i : y_i = +1; \\ -1, if \exists y_i : y_i = -1 \end{cases}$$

Based on this, the MIL assumption is widely adopted in binary classification problems, where a positive packet contains at least one positive instance, but a negative packet contains only negative instances. Thus, the homogeneity of data in the negative packet is the key feature of the method, which enables efficient learning.
The paper introduced a novel MIL model in order to predict sample-level purity from H

E stained digital histopathology slides with three core procedures, which are: a feature extractor module: Neutral Network, a MIL 'distribution' pooling filter, and a bag-level representation transformation module, Neutral Network. The feature extractor module is trained to extract features from plaque tumors and normal sections of the patient. A series of segmentation maps are obtained by hierarchical clustering of the extracted feature vectors.

4 Results

Based on a series of conclusions given in the paper, the reasons for the success of the MIL model were first analyzed from a statistical perspective: by combining correlation analyses, using Spearman's correlation coefficients, based on Fisher transformation constructs and comparing the confidence interval of as a basis for judgment, high Spearman coefficients therefore illustrate a significant correlation between tumor purity and genomic tumor purity values was obtained, followed by the conclusion that the MIL model has a lower mean absolute error compared to the tumor nuclei estimation method from the perspective of standard deviation.
Secondly, from the degree of implementation of the MIL model, the feature extractor can not only learn discriminant features for cancer cells and normal cells, but also cluster the feature vectors to obtain discriminant features. In addition, the ROC curve analysis based on the percentile bootstrap method was used for each cohort, and satisfactory results were obtained in the case of samples with normal slides.
From the spatial perspective of tumor purity prediction, due to the nature of Intra-tumor heterogeneity, the investigators tested both the top and bottom slides of a sample, randomly selected from 100 bags as model training inputs to obtain the predicted values, and used a non-parametric Wilcoxon signed-rank test for both slides in order to obtain the statistical hypotheses and results, and by comparing the true absolute error and expected absolute error, using Wilcoxon signed-rank test, finally determining that the prediction results of slides with both top and bottom are more accurate.

Conclusions

This article cleverly combines deep learning models and tumor prediction, making a series of detailed and innovative assumptions in order to reduce the computational workload of pathologists and improve the accuracy of machine learning models in predicting tumor purity under different conditions and in different ways. Although the MIL model is highly accurate both in terms of model performance and regression effects, this experiment locks the direction of the choice of samples to the two external cohorts, but if more kinds of preservation methods can be considered, the probability of tissue contamination can be circumvented as much as possible, and the systematic error can be reduced nearly, thus obtaining a more unbiased model.

References

[1] Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, Daniel Shao Weng Tan, Weiwei Zhai, Anders Jacobsen Skanderup, Wing-Kin Sung, and Hwee Kuan Lee. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. *bioRxiv*, 2021.

[2] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

[3] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018.

[4] Dan Guo, Melanie Christine Föll, Veronika Volkmann, Kathrin Enderle-Ammour, Peter Bronsert, Oliver Schilling, and Olga Vitek. Deep multiple instance learning classifies subtissue locations in mass spectrometry images from tissue-level annotations. *Bioinformatics*, $36(\text{Supplement}_1) : i300 - -i308, 072020$.